Traffic Data Classification using Machine Learning Algorithms in SDN Networks

Jungmin Kwon*, Daeun Jung*‡ and Hyunggon Park*†‡
*Dept. Electronic and Electrical Engineering, Ewha Womans University, Seoul, Republic of Korea
‡Smart Factory Multidisciplinary Program, Ewha Womans University, Seoul, Republic of Korea
†The Alan Turing Institute, London, United Kingdom
jungmin.kwon@ewhain.net, daeun.jung@ewhain.net, hyunggon.park@ewha.ac.kr

***Abstract*—As an efficient approach to proactively monitoring network dynamics, automatically analyzing network data, and predicting network usage, machine learning has been widely deployed. This enables the networks to be efficiently and autonomously coped with in SDN/NFV environment. In particular, network intelligent technology can be adopted into the infrastructure management, network operations, and service assurance. In this paper, we study the automatic network data classification based on machine learning, where several machine learning algorithms are deployed to automatically classify real network traffic data collected from ONOS (Open Network Operating System) platform. From the experiment results with simple network topology, we conclude that machine learning algorithms can effectively classify the network traffic data. However, it is also observed machine algorithms may only show a limited performance in practice if they are blindly deployed. This is because there exists not only the data that needs to be delivered to the receivers but also the data required for network maintenance in a real network system. Therefore, it is essential to develop machine learning algorithms that explicitly consider the characteristics of real network traffic data in target network scenarios.**

***Index Terms*—Machine learning, supervised learning, automatic network data classification, ONOS**

I. Introduction

With the advance of 5G and Beyond 5G (B5G) mobile and wireless networks, network automation becomes paramount. The network automation is known as the process of automating the network configuration, network management, network deployment, and network operation of physical and virtual resources and devices. The benefits for deploying network automation are not only the efficiency and robustness of the network maintenance but also the cost reduction associated with network operations. This becomes more important as a large number of devices are simultaneously connected and supported by a variety of services.

Machine learning has been deployed in networks as an efficient approach to proactively monitoring network dynamics, automatically analyzing network data, and predicting network usage. This enables the networks to be efficiently and autonomously coped with in SDN/NFV environment. It is claimed from the standardization group, ETSI ISG (industry specification group) ENI (experiential network intelligence), that network intelligent technology can be adopted into the infrastructure management, network operations, and service assurance [1]. The infrastructure management involves the prediction of traffic flows [2], network resource management [3]–[5], network auto-scaling [6], [7], and optimal tracking control [8]. The network operations include abnormal detection [9], [10], quality of decision (QoD) improvement and network monitoring cost minimization [11], and VNF failure prediction [12]. The service assurance contains application identification [13], [14], VNF placement and chaining [15], [16] and content popularity prediction [17].

One of the basic functionalities of network automation is the network data analytics where it analyzes and predicts real-time network traffics, so that the networks can be efficiently managed. The network traffic data generated from various types of network protocols and users in both physical and virtual networks has often been manually analyzed. For efficient network data analytics, however, machine learning algorithms can be adopted [18], while it is still important issues how to identify appropriate algorithms and how to efficiently use them for network data analytics.

In this paper, we focus on the network data analytics based on machine learning algorithms with the actual data collected from an ONOS SDN controller, which can be included in 5G core networks. We study how machine learning algorithms can be adopted in the SDN/NFV environment for network traffic data analytics. While it is shown from the experiment results that the machine learning algorithms can be used for network traffic data analytics, blindly deploying machine learning algorithms in network traffic classification does not guarantee good performance in practice. Rather, it is essential to develop machine learning algorithms dedicated to network traffic data classification with the understanding of network protocols and target network scenarios.

This paper is organized as follows. In Section II, we describe the actual implementation of system architecture for network traffic data collection and analysis. We then describe the detailed information about the data features used for the

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00024, Supervised Agile Machine Learning Techniques for Network Automation based on Network Data Analytics Function) and supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2B5B01002528).

TABLE I
LIST OF NETWORK TRAFFIC DATA FEATURES

Feature	Description
time	Data collection time
bytesReceived	Cumulative size of received packets
bytesSent	Cumulative size of sent packets
durationSec	Time of port activation
packetRxDropped	Information of dropped packets at the receiver node
packetsRxError	Information of packet errors at the receiver node
packetsSent	The number of sent packets
packetsTxDropped	Information of dropped packets at the sender node
packetsTxError	Information of packet errors at the sender node
packetsReceived	The number of received packets
port	Information of the port
rx_throughput	Throughput at the receiver node
tx_throughput	Throughput at the sender node

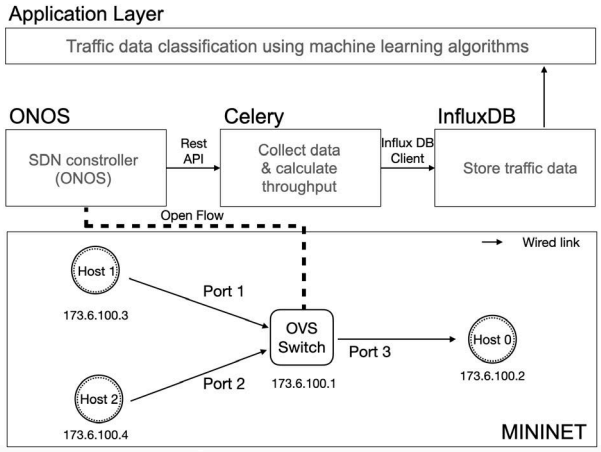

Fig. 1. An illustrative description of experiment network model

machine learning algorithms. Experiment results are presented in Section III and the conclusions are drawn in Section IV.

II. SYSTEM SETUP

In this section, we present how machine learning algorithms can be adapted to classify network traffic data that are collected from real network implementation.

The real network traffic data collected by REST API from ONOS controllers consist of various network features including the port number as shown in Table I. In order to identify various patterns of network traffic, we primarily classify the sender node using traffic data based on a supervised learning algorithm such that the ONOS controller can identify the sender's port number.

We consider a simple network topology using Mininet based simulation framework, where it consists of two sender nodes ('Host 1' and 'Host 2'), OVS switch, and receiver node ('Host 0'), as shown in Fig. 1. These network packets are stored at InfluxDB in the shape of traffic information, which is calculated from the task queue manager called Celery. The network traffic data is used for training and testing machine learning algorithms at the application layer.

In this paper, we adopt three different machine learning algorithms, Random Forest (RF), Linear Discriminant Analysis (LDA), and DNN, for network traffic data classification. Since the port number information connected to each host is unique, as shown in Fig 1, we classify three different sender nodes 'Host 1', 'Host 2', and OVS using the port numbers. Note that the classification of traffic data port numbers can be considered as a sender node classification since there is one-to-one mappings between port numbers and the sender nodes.

III. EXPERIMENT RESULTS

A. Experiment Setup

In order to evaluate the performance of machine learning algorithms over practical physical and virtual networks, we consider two different scenarios, a regular data delivery over the network (Scenario A) and a malicious network that sporadically attacks the receiver node by flooding too many requests (Scenario B). In both scenarios, the traffic data generated from sender nodes is PING request messages. Moreover, we assume that the sender nodes transmit the network packet in a certain duration followed by the Poisson distribution. However, in Scenario B, sender node 'Host 1' generates PING request messages in time intervals, while 'Host 2' attacks the receiver node at rare intervals by PING-flooding.

The links between each node and switch are set with a default configuration in ONOS platform. Therefore, there is no packet drop at every link, as the network is implemented in wired networks. While both sender nodes transmit the data with the same time durations in Scenario A, two sender nodes transmit the data with different transmission patterns in Scenario B as we assume that there is an abnormal node in this scenario. 'Host 2' which is set as an abnormal sender node in Scenario B sporadically transmits 200 times more data than 'Host 1' for a short time when it attacks the receiver node.

B. Evaluation

In this section, we evaluate the performance of machine learning algorithms with the network traffic data collected from the system that we implemented. We collected 53,171 traffic data from Scenario A and 2,036 traffic data from Scenario B in ONOS controller. In order to train the machine learning model, 70% of data samples are used for the train set and the rest of the data for the test set.

Table II shows the performance of sender node classification in both scenarios. In scenario A, the classification accuracies of RF and LDA are 95% and 98%, respectively, while DNN shows 69% accuracy. Although the DNN performance is lower than other algorithms, it is improved from 47% accuracy in [19], where the DNN structure was not optimized for network traffic data in [19].

In Scenario B, on the other hand, it is observed that the performance of abnormal behavior detection with the same algorithms is significantly degraded, where the accuracies of RF, LDA, and DNN and RF are 42%, 76%, and 74%, respectively. Specifically, RF shows lower than 50% accuracy since

TABLE II
SENDER NODE CLASSIFICATION ACCURACY IN SCENARIO A AND B

		RF	LDA	DNN
Accuracy δ (%)	Scenario A	95%	98%	69%
	Scenario B	42%	76%	74%

the traffic data collected from an abnormal scenario incurs a large variations in the structure of the optimal decision tree in RF. In case of DNN, we also compare the DNN performance with blindly used DNN and found that the accuracy of the DNN in this paper is 74% which is improved compared to the performance in [20]. The reason why the performance degradation of the DNN algorithm is because the network data for analysis includes the unintended data generated by various network protocols such as ARQ messages. Hence, these experiments show that deploying machine learning algorithms in traffic data classification, without understanding the detail of network protocols and target scenarios, does not guarantee good performance in real physical and virtual networks.

IV. CONCLUSION

In this paper, we present preliminary results for automatic network data classification based on machine learning algorithms. One of the challenges for real network data analytics based on machine learning algorithms is from the inclusion of unintended data generated for network management. In order to analyze the traffic data collected from the SDN/NFV environment, we deploy several machine learning algorithms for identifying network traffic data. We illustrate how machine learning algorithms can be adopted for network data analytics by using actual network traffic data collected from the ONOS SDN controller in two different target network scenarios. While it is shown from the experiment results that the machine learning algorithms can be used for network traffic data analytics, blindly deploying machine learning algorithms in network traffic classification does not guarantee good performance in practice. Hence, it is important to develop machine learning algorithms for network data analytics, as future works.

REFERENCES

[1] H. Kim, M. Shin, B. Ahn, J. Lee, S. Lee, S. Lee, J. Ham, and S. Hyeon, "Network intelligence technologies," *ETRI Insight*, 2018.

[2] W. Jiang, M. Strufe, and H. D. Schotten, "Experimental results for artificial intelligence-based self-organized 5G networks," in *IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.

[3] S. Sun, L. Gong, B. Rong, and K. Lu, "An intelligent SDN framework for 5G heterogeneous networks," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 142–147, 2015.

[4] A. Martin, J. Egaña, J. Flórez, J. Montalbán, I. G. Olaizola, M. Quartulli, R. Viola, and M. Zorrilla, "Network resource allocation system for QoE-aware delivery of media services in 5G networks," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 561–574, 2018.

[5] Y. He, F. R. Yu, N. Zhao, V. C. M. Leung, and H. Yin, "Software-Defined networks with mobile edge computing and caching for Smart cities: A big data deep reinforcement learning approach," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 31–37, 2017.

[6] S. Rahman, T. Ahmed, M. Huynh, M. Tornatore, and B. Mukherjee, "Auto-scaling VNFs using machine learning to improve QoS and reduce cost," in *IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[7] P. Tang, F. Li, W. Zhou, W. Hu, and L. Yang, "Efficient Auto-Scaling approach in the telco cloud using self-learning algorithm," in *IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.

[8] J. Leguay, L. Maggi, M. Draief, S. Paris, and S. Chouvardas, "Admission control with online algorithms in SDN," in *IEEE Network Operations and Management Symposium (NOMS)*, 2016, pp. 718–721.

[9] G. A. Ajaeiya, N. Adalian, I. H. Elhajj, A. Kayssi, and A. Chehab, "Flow-based intrusion detection system for SDN," in *IEEE Symposium on Computers and Communications (ISCC)*, 2017, pp. 787–793.

[10] L. Fernández Maimó, A. L. Perales Gómez, F. J. García Clemente, M. Gil Pérez, and G. Martínez Pérez, "A self-adaptive deep learning-based system for anomaly detection in 5G networks," *IEEE Access*, vol. 6, pp. 7700–7712, 2018.

[11] V. Sciancalepore, F. Z. Yousaf, and X. Costa Perez, "z-TORCH: An automated NFV orchestration and monitoring solution," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1292–1306, 2018.

[12] H. Huang and S. Guo, "Proactive failure recovery for NFV in distributed edge computing," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 131–137, 2019.

[13] Z. A. Qazi, J. Lee, T. Jin, G. Bellala, M. Arndt, and G. Noubir, "Application-awareness in SDN," in *Proceedings of the Association for Computing Machineryz SIGCOMM*, vol. 43, no. 4, 2013, pp. 487– 488.

[14] P. Amaral, J. Dinis, P. Pinto, L. Bernardo, J. Tavares, and H. S. Mamede, "Machine learning in software Defined Networks: Data collection and traffic classification," in *IEEE International Conference on Network Protocols (ICNP)*, 2016, pp. 1–5.

[15] J. Pei, P. Hong, and D. Li, "Virtual Network Function selection and chaining based on deep learning in SDN and NFV-enabled networks," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, 2018, pp. 1–6.

[16] X. Zhang, C. Wu, Z. Li, and F. C. M. Lau, "Proactive VNF provisioning with multi-timescale cloud resources: Fusing online learning and online optimization," in *IEEE Conference on Computer Communications (INFOCOM)*, 2017, pp. 1–9.

[17] W. Liu, J. Zhang, Z. Liang, L. Peng, and J. Cai, "Content popularity prediction and caching for ICN: A deep learning approach with SDN," *IEEE Access*, vol. 6, pp. 5075–5089, 2018.

[18] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE access*, vol. 6, pp. 32 328–32 338, 2018.

[19] J. Kwon, D. Jung, and H. Park, "Study on network data feature importance for network node classification," in *The Korean Institute of Communications and Information Sciences (KICS) Winter Conference*, Jan 2020, pp. 1245–1246.

[20] ——, "Study on machine learning based abnormal behavior network node classification," in *The Korean Institute of Communications and Information Sciences (KICS) Winter Conference*, Jan 2020, pp. 1232–1233.